3.2.6 Gevoeligheid aan outliers

Om de gevoeligheid aan outliers van de verschillende spreidingsmaten te onderzoeken, bekijken we opnieuw het verschil in reactietijd tussen de congruente en incongruente opdrachten van de blanken, zoals weergegeven in Figuur 3.5 op pagina 83.

We berekenen eerst de spreidingsmaten op basis van alle waarden (inclusief de outlier), en herhalen dit dan voor de waarden zonder de outlier. Indien er een groot verschil is tussen het resultaat met of zonder de outlier, besluiten we dat de maat gevoelig is aan outliers.

Tabel 3.6 geeft deze berekeningen weer en illustreert dat de variatiebreedte v_X, de gemiddelde absolute afwijking ga_X, de variantie s_X^2 en de standaarddeviatie s_X gevoelig zijn aan outliers. Van deze spreidingsmaten zijn voornamelijk de variatiebreedte en variantie zeer gevoelig aan outliers.

De interkwartielafstand Q anderzijds is niet gevoelig aan outliers.

	v_X	ga_X	s_X^2	s_X	Q
met outlier	2492.40	181.90	108724.76	329.73	201.46
zonder outlier	974.82	150.13	41527.03	203.78	199.79

Tabel 3.6: De spreidingsmaten voor het verschil in reactietijd bij de blanken op basis van alle waarden (inclusief de outlier) en op basis van de waarden zonder de outlier.

De spreidingsmaat d is vooral nuttig voor nominale en ordinale variabelen. We zullen dit niet berekenen voor het verschil in reactietijd omdat elke blanke persoon een unieke waarde heeft, waardoor d dicht bij 1 zal liggen.

Omdat de spreidingsmaat d afhangt van de frequentie van de modus, het aantal unieke waarden en de steekproefgrootte, is ze niet gevoelig aan outliers.

3.3 Boxplot

Op basis van de kwartielen en de interkwartielafstand, zoals besproken op pagina 92, kunnen we een nieuwe figuur maken: de *boxplot*. Een boxplot kan opgesteld worden zonder data te groeperen en is bijgevolg niet gebruikersafhankelijk. Dit is verschillend van het histogram.

Hoewel een boxplot er op het eerste zicht wat vreemd uitziet, is het een zeer bruikbare

figuur eenmaal je ze gewoon bent. Het zal ons in staat stellen om een idee te krijgen over de verdeling van de data en om *outliers* visueel vast te stellen. We gebruiken de volgende rekenregel om te bepalen of een waarde van een variabele een outlier is. We berekenen eerst de interkwartielafstand Q en vervolgens het verschil:

$$P_{25} - 1.5 \times Q,$$

dus het eerste kwartiel min 1.5 keer de interkwartielafstand. Alle waarden die *kleiner* zijn dan dit verschil, zijn outliers (volgens deze rekenregel). Vervolgens berekenen we de som:

$$P_{75} + 1.5 \times Q,$$

dus het derde kwartiel plus 1.5 keer de interkwartielafstand. Alle waarden die *groter* zijn dan deze som, zijn ook outliers. Outliers kunnen dus zowel grote als kleine waarden zijn. Het is uiteraard ook mogelijk dat een variabele *geen* outliers heeft.

In Figuur 3.9 construeren we stap voor stap een boxplot op basis van de variabele Leeftijd zoals gegeven in Tabel 2.6 op pagina 45. We starten door een verticale as te tekenen die de leeftijd voorstelt. Vervolgens zetten we naast de as een stip voor de leeftijd van elk van de 90 personen in de steekproef (Figuur A). De eerste persoon in Tabel 2.6 is 40 jaar, dus zetten we een stip ter hoogte van 40 naast de verticale as. De tweede persoon is 21, dus zetten we een stip ter hoogte van 21 naast de verticale as, etc.

Vervolgens gebruiken we de rekenregels om outliers te bepalen. Voor Leeftijd is het eerste kwartiel $P_{25} = 21$, het derde kwartiel $P_{75} = 37$ en de interkwartielafstand $Q = 37 - 21 = 16$. Bijgevolg is $P_{25} - 1.5 \times Q = 21 - 1.5 \times 16 = -3$. Dus alle personen jonger dan -3 jaar zijn outliers volgens de rekenregel. Het is evident dat er geen personen zijn met een negatieve leeftijd, dus zijn er geen outliers bij de kleine waarden. Voor de grote waarden berekenen we $P_{75} + 1.5 \times Q = 37 + 1.5 \times 16 = 61$. Alle personen die ouder zijn dan 61 jaar, zijn dus outliers. Als we kijken naar Tabel 2.6 zien we dat er 4 personen ouder zijn dan 61 jaar. In de Figuur 3.9 zullen we de stippen horende bij deze leeftijden rood kleuren (Figuur B).

In een volgende stap (Figuur C) tekenen we een horizontale lijn[m] bij de laagste stip die niet rood is (dus de kleinste waarde die geen outlier is). Aangezien er geen outliers zijn bij de kleine waarden, komt dit overeen met de minimumleeftijd (hier 16 jaar). We tekenen ook een horizontale lijn bij de hoogste stip die niet rood is (dus de grootste die geen outlier is).

Voor de kwartielen doen we iets gelijkaardigs: we tekenen een horizontale lijn ter hoogte van het eerste kwartiel $P_{25} = 21$, en ter hoogte van het derde kwartiel $P_{75} = 37$ (Figuur

[m] De breedte van de horizontale lijn mag je zelf kiezen.

D). De horizontale lijnen ter hoogte van de kwartielen verbinden we nu met elkaar met verticale lijnen zodat we een rechthoek bekomen (Figuur E).

In een volgende stap (Figuur F) verwijderen we alle stippen die niet rood zijn (dus alle waarden die geen outliers zijn). In een voorlaatste stap (Figuur G) tekenen we een verticale stippellijn van de onderste horizontale lijn tot het eerste kwartiel, en van het derde kwartiel tot de bovenste horizontale lijn. Deze verticale stippellijnen worden ook de *whiskers* of *snorharen* genoemd.

De laatste stap bestaat uit het tekenen van een horizontale lijn in de rechthoek ter hoogte van de mediaan $P_{50} = md_X = 26$. Deze laatste figuur stelt de boxplot voor van de variabele Leeftijd.

Een boxplot is handig omdat het visueel volgende informatie bevat:

- de mediaan (centrummaat): de horizontale lijn in de rechthoek.
- de interkwartielafstand (spreidingsmaat): de hoogte van de rechthoek.
- de outliers: de observaties die door bolletjes zijn aangeduid.

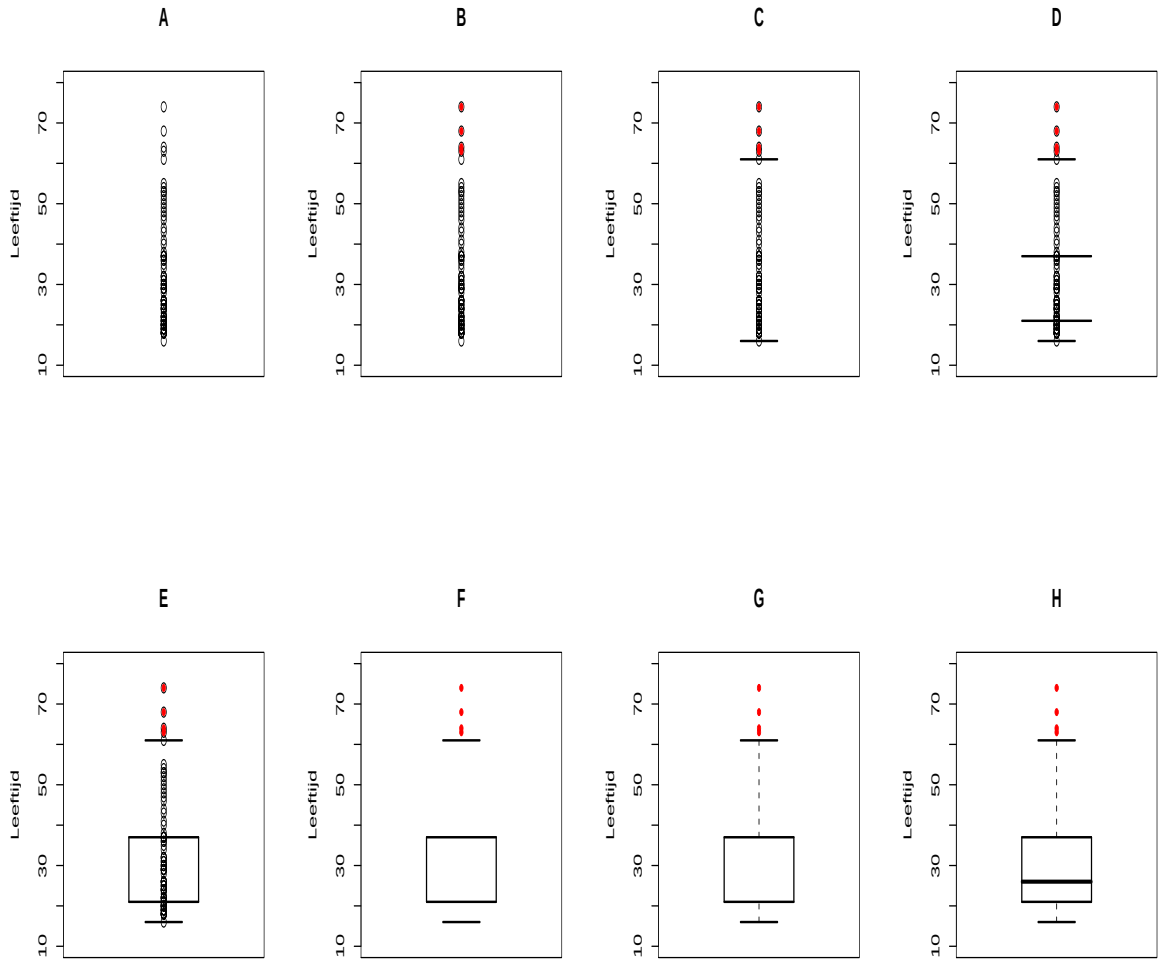
Figuur 3.10 toont de boxplots van Leeftijd in de 3 hypothetische steekproeven uit Figuur 2.10 op pagina 50. De kleine cirkels duiden hier de outliers aan.

Het is ook mogelijk om een boxplot *horizontaal* te tekenen i.p.v. *verticaal*. Figuur 3.11 geeft de horizontale variant van Figuur 3.10 die ontstaat door de verticale boxplots een kwartslag te draaien. Voor de scheve verdeling naar rechts (Figuur 3.11 links) zien we dat er redelijk wat outliers rechts zijn. Dit houdt steek: er zijn vooral jongeren in deze steekproef, dus de enkele personen die ouder zijn, worden als outliers beschouwd. Bij de scheve verdeling naar links is het net omgekeerd, terwijl er bij de symmetrische verdeling zowel enkele jongeren als ouderen outliers zijn.

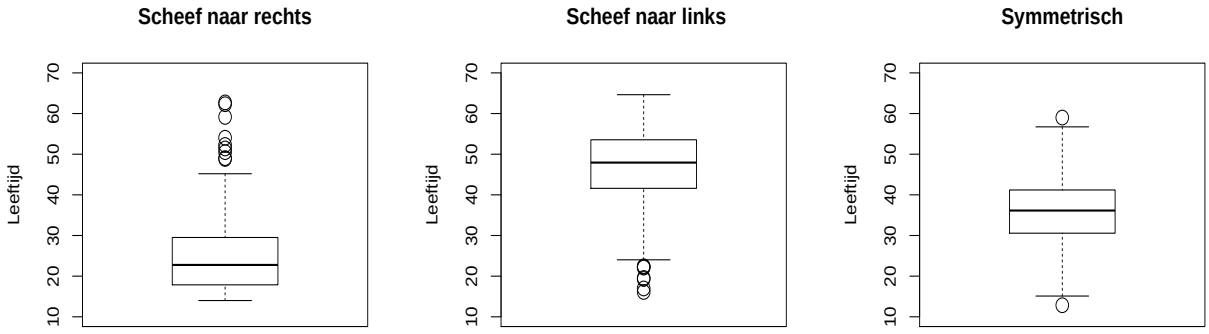
Illustratie in R

Via `boxplot()` bekomen we een boxplot in R.

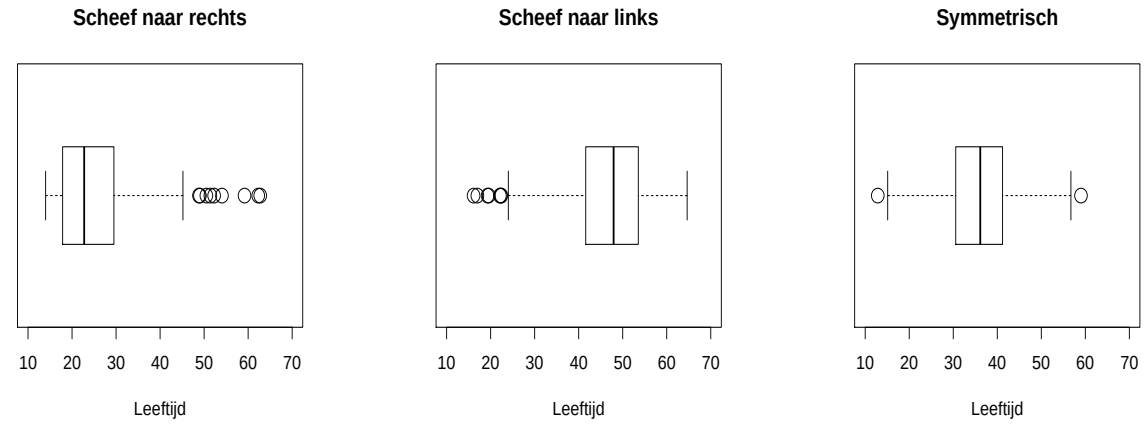
```
> boxplot(DataIAT$Leeftijd)
```



Figuur 3.9: Constructie van een boxplot voor de variabele Leeftijd. We verwijzen naar de tekst voor de uitleg.



Figuur 3.10: De boxplots van Leeftijd voor 3 verschillende (hypothetische) steekproeven.



Figuur 3.11: De horizontale weergave van de boxplots voor 3 verschillende (hypothetische) steekproeven.